Neuroimaging Data Landscapes

Annex to SCARP Case Study No. 1

Angus Whyte
Digital Curation Centre, University of Edinburgh

Deliverable B4.8.2.1 Annex

Version No.	1.1
Status	**FINAL**
Date	11 November 2008

Copyright

Catalogue Entry

Title Neuroimaging Data Landscapes: Annex to SCARP Case Study 3
Creator Angus Whyte (author)
Subject Data curation; formats, processes and issues; system development; standards; legal factors; methodology, and problems overcome; human factors
Description This is the Annex to Case Study No. 1 of the Digital Curation Centre's SCARP Project titled 'Curating Brain Images in a Psychiatric Research Group: Infrastructure and Preservation Issues' (SCARP Deliverable number B4.8.2.1). It comprises a literature review and discussion of the study methodology, which elaborate on the main report's treatment of these.
Publisher University of Edinburgh; UKOLN, University of Bath; HATII, University of Glasgow; Science and Technology Facilities Council.
Date 11 November 2008 (creation)
Type Text
Format Adobe Portable Document Format v.1.3
Resource Identifier ISSN 1759-586X
Language English
Rights © 2008 DCC, University of Edinburgh

Citation Guidelines

Whyte, A. (2008), " Neuroimaging Data Landscapes: Annex to SCARP Case Study No. 1", Digital Curation Centre, Retrieved <date>, from http://www.dcc.ac.uk/scarp

Contents

Introduction

This report '*Neuroimaging Data Landscapes*' is an Annex to Case Study No. 1 of the Digital Curation Centre's SCARP Project titled '*Curating Brain Images in a Psychiatric Research Group: Infrastructure and Preservation Issues'* (SCARP Deliverable number B4.8.2.1), or 'the main report' as it is referred to below. It comprises a literature review and discussion of the study methodology, which elaborate on the main report's treatment of these.

1. "Little Big Science"- Neuroimaging in Psychiatry

1.1 Introduction

This first chapter surveys the 'landscape' of neuroimaging; first giving some background to the case and, in the next section below, on the development of imaging as a technique for psychiatric research. This introduces the nature of imaging studies, the range of primary data collected, and the techniques used to derive analytic data. Then section 1.3 considers the rationale for sharing data in the neuroimaging field, the development of policy in that area by research councils and professional bodies, and the legal and ethical constraints on sharing. Section 1.4 looks at the availability of neuroimaging data and publication repositories, the emerging infrastructure for data sharing, integration and reuse; and the range of considerations and models in use for providing collaborators and independent researchers with access to research data.

Neuroimaging has been characterised as a 'little big science' (Beaulieu, 2002). On the one hand it is 'little' in that the traditional disciplines that make use of it- psychiatry in the present case, and psychology and neurology amongst others - are not highly technologically dependent fields of research. Technological dependence is, on the other hand characteristic of neuroimaging, reliant as it is on magnetic resonance image (MRI) scanners as a data source, with a typical capital cost of £1.5 million, and on computationally intensive means of analysis to provide its results. This reliance on capital-intensive instrumentation is a characteristic of 'big science' and increasingly of psychiatric neuroimaging. For reasons discussed in this chapter, neuroimaging is in some ways archetypal of fields that are embracing e-Science (or e-Research); with considerable investment in national and international collaborative efforts to share data and harmonise data analysis techniques and terminology.

1.2 Imaging and the Neurological approach to Psychiatric Disorders

A helpful introduction to the growth in 'size' of neuroimaging towards big science comes in this market research report excerpt from US firm Frost and Sullivan:-

> "Magnetic resonance imaging rightly holds its place near the top of the medical imaging modality value chain. It offers exceptional quality in terms of resolution, particularly in brain imaging, and its use of a strong magnetic field, as opposed to ionising radiation, eliminates a major safety concern for patients. Such imaging advantages come at a cost, however. The modality hardware itself for high end equipment comfortably exceeds $ 1 million. With this capital expenditure comes an additional expenditure of energy; the headache of correctly siting a scanner so that the hospital /radiology practice can accommodate the equipment's strong magnetic field. This is proving an extra strain on the hard-stretched budgets of hospitals across Europe, and is a significant barrier to the acquisition of a new MRI unit, or - increasingly so with the advent of 3 Tesla machines – upgrading to a new one. Despite these provisos, the stringent following of siting and safety guidelines can ensure that the maximum potential of an MRI scanner is realised, both in terms of image quality and patient throughput." (Bryant, 2005)

As this suggests, MRI has become part of the infrastructure of clinical practice, but with certain strings attached; location, safety, and large capital budgets among them, and the skills of radiology units and engineers to extract high quality images of (for example) the brain. MRI is a relatively recent form of imaging however, and among other advantages for brain research are that it provides a non-invasive technique for researching what is

'in' the brain and, with the development of functional imaging (fMRI) techniques, what is 'going on' in the brain. This has made it a particularly attractive technique for research that aims to find neurobiological explanations for psychiatric disorder, since imaging can depict changes in and between individual brains, and correlate those changes with a range of behavioural, social and clinical phenomena.

Before considering the kinds of brain changes sought in imaging research and their relevance to diagnosis, some historical background is worthwhile. MRI technology developed from experiments in 1946 by US physicists Bloch and Purcell. They demonstrated the phenomenon of *nuclear magnetic resonance* (NMR) by showing that when certain nuclei were placed in a magnetic field they absorbed energy in the radiofrequency range of the electromagnetic spectrum, and that this energy could be detected as the nuclei transferred to their original state (Ellard, 2007). This discovery was applied to biological samples in the 1960s and in the 1970s it was noticed that NMR signals changed in diseased tissue (Pekar, 2006).

It was also in the 1970's that NMR signals were first used to form images, the basis for MRI. NMR is able to provide useful information on the composition of magnetised material because the strength of the radio frequency signal depends on several factors; the varying density of protons in the atoms comprising biological tissue, which oscillate with a frequency dependent on the strength of the magnetic field they are exposed to, and the time the protons take to 'relax' back into an equilibrium state. The NMR also varies with the differences in the parallel and perpendicular 'relaxation time' of magnetised tissue relative to the magnetic field. In 1973 Lauterbur applied a gradient of the magnetic field to make the resonance frequency of nuclei vary linearly with their spatial location, demonstrating that this allowed a spatial representation of the frequencies and their variations with different kinds of tissue; making it possible to construct an *image* of these variations (ibid.)

Compared with most soft tissues of the body the density of different kinds of brain tissue varies considerably. When the use of MRI for clinical purposes took off in the late 1980s and early 1990s with the commercial availability of scanners, it therefore proved particularly useful to research in stroke and brain tumours (Pekar, 2006.). It also enabled research on changes in brain structure that could similarly demonstrate an organic basis for psychiatric disorders like schizophrenia (Lawrie et al, 2005).

MRI scan data is first saved as two-dimensional 'slices', using a mathematical formalism known as '*k*-space' to encode the scan signal phase and frequency as a representation of image density. It is by applying a second formalism - a Fourier transform - that the encoded signal is transformed into a three-dimensional representation of the brain as a volume, and this is termed *reconstruction.* (Johnson and Becker, 1999, Pekar, 2006).

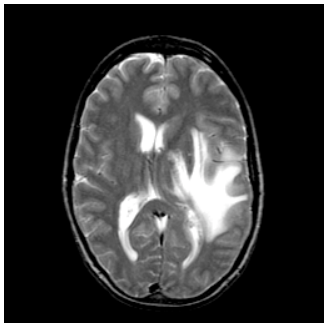


Figure 1.1: MRI slice (source: *Whole Brain Atlas*, Johnson and Becker, 1999)

The images from MRI are snapshots depicting densities as grey-scale values (1 to 256) at each of a number of 3-dimensional pixels or *voxels*. Each snapshot is a 'slice', typically 3 to 5mm thick and normally oriented across a vertical axis so that the slice is displayed as if the viewer is looking up through the brain (Johnson and Becker, 1999). The resolution of the image depends on several factors; the 'flip angle' of the protons in their magnetised state, and the cycle time between radio-frequency pulses used to obtain the signal. Volumetric measurements are possible by multiplying the slice thickness by the area of interest. Since a sequence of slices is taken in a scanning session it is possible to identify the volumes occupied by specific anatomical regions

of the brain, such as the hippocampus and the amygdyla, by mapping these to a *template* conforming to known anatomical regions, derived from a brain atlas (Lawrie et al, 2005).

Psychiatric research using *structural* MRI has focused on changes in specific brain regions. Analyses may be of differences at one point in time between groups of patient and control brains, or sometimes changes in brains between scanning sessions over months or years. Using the earlier scanning technology computed tomography (CT), in 1976 the Edinburgh group was the first to report associations between such changes and impairment in schizophrenic patients (Johnstone et al 2003). The greater resolution and differentiation of brain regions possible with MRI has contributed a great deal to further study by it and other centres of imaging research (Lymer et al 2006) and it is now feasible to predict the development of psychotic symptoms in young people at genetic risk of schizophrenia, from changes in the density of particular brain regions (Job et al, 2007)

Structural imaging studies allow changes in individuals' brains to be studied. *Functional* imaging (fMRI) is a relatively recent development enabling brain *processes* to be represented from scanning data. In 1990 it was demonstrated that the appearance of the brain's blood vessels changed with blood oxygenation, and this rapidly led to use of blood-oxygenation-level-dependent (BOLD) contrast to study brain *activation* in anatomical regions (Pekar, 2006).

fMRI studies measure brain activity indirectly, in terms of the flow and oxygenation level of blood in the brain. fMRI experiments study how these change in response to task 'stimuli' that participants (or subjects) are asked to respond to inside the scanner. fMRI scanning captures slice images rapidly, sacrificing the image resolution of structural scanning in favour of the added dimension of time, and building up a movie-like sequence. Using that analogy each 'frame' corresponds to a brain volume of (e.g.) 30-40 slices captured within one 'time of repetition' (TR). Pekar (2006) describes a typical fMRI session as follows:-

> "For a typical TR of 2 s, if 200 volumes are acquired, then we have a volume movie consisting of 200 volumes of the brain (each consisting of 30–40 slices) acquired over 400 s. During these 400 s, a neuro-behavioral paradigm is played out in which the research participant is exposed to sensory stimuli or asked to perform some set of mental and motor tasks or some combination of them. So we have a situation where 400 s of temporally structured brain activity (e.g., watching flashing lights every other 30 s, tapping one's fingers every other 20 s, reading words, or solving math problems) are accompanied by the acquisition of a brain volume movie with 2 s temporal resolution" (Pekar, 2006, p.25)

fMRI is now used very widely in neuroscience research. In psychiatry it has been used for example to study relationships between individuals' brain activity when completing sentences and classifying words, and their assessed ratings on various psychotic symptoms, providing insights into areas of the brain associated with hallucinations and delusions (Whalley et al 2007).

Neuroimaging Study Design

Imaging studies normally follow the case-control design that predominates in medicine. A sample of cases is compared with unaffected control groups, and possibly others including healthy relatives - or individuals with another psychiatric disorder. The general aim is to test hypothesised differences between groups, for example in the size or density of brain regions (Lawrie, op. cit.). Typical study stages are shown in Figure 1.2

below. This section gives an overview of the design to analysis stages, returning to publication and archiving later.

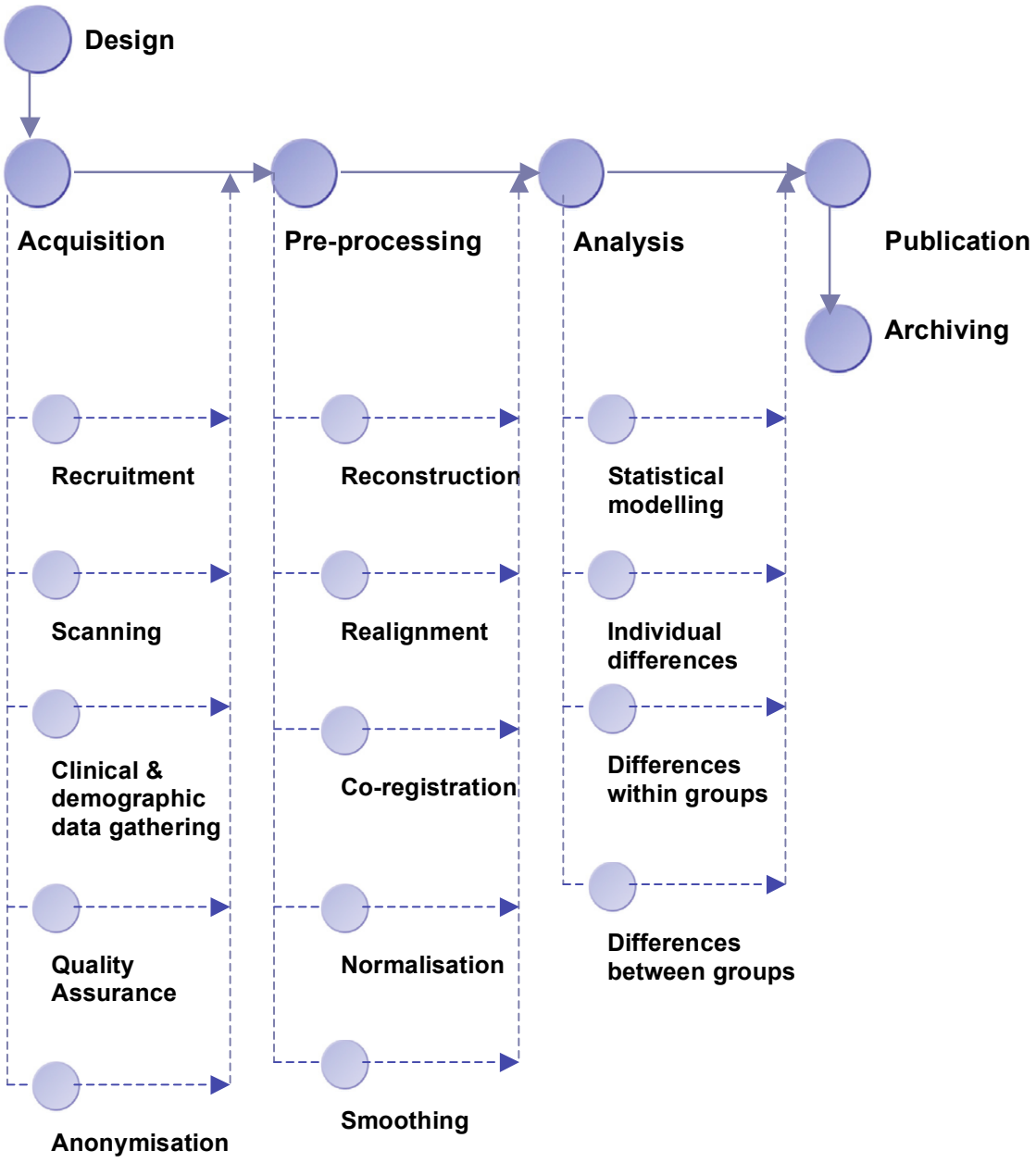


Figure 1.2 Typical stages of a functional neuroimaging study

Studies normally test a hypothesis defined at the *design stage*, and involving a model of the explanatory variables; for example demographic and neuropsychiatric variables as well as (in functional studies) the key aspects of whatever 'stimulus' the subject is asked to respond to, and also 'noise factors' such as head movements or the heart beating.

Structural MRI studies are *observational*, being concerned with capturing individuals' brain anatomy at particular points in time. Functional MRI studies on the other hand have an e*xperimental* design, since the brain functions are analysed in relation to hypothesised effects of a task stimulus, and design affects all the subsequent analysis

steps. According to Miyapuram et al (2007), three main types and two classes of design are commonly identified: categorical, factorial and parametric design types differ in their assumptions about brain process; whereas epoch and block-related design classes differ in task-setup.

At the design stage *ethical approval* must be obtained from the relevant NHS Research Ethics Committee, to carry out the research while safeguarding the interests of the human subjects it intends to involve. The subjects – whether patients, healthy relatives or control groups - must of course also be recruited and their informed consent obtained to take part in the study. These legal and ethical requirements, which the next section returns to in more detail, mean that studies have to be designed to a high degree before any data is actually acquired.

Data Acquisition

Neuroimaging is used in a wide variety of neuroscience research involving humans and other animals. In psychiatry in particular, MRI scan data is of little use without detailed information about the person scanned. Since neuroimaging studies aim to find variables that explain differences in their brain structures and activity, a great deal of quite complex data is gathered on the research subjects or participants (Keator et al, 2006) The data gathered includes *demographic* data – typically that known (or assumed) to have some bearing on brain size and physiology, such as age, sex, height, and handedness. Also gathered are data with some known relation to mental health; for example the Edinburgh High Risk Study (Johnstone et al 2003) includes social and economic classification data, information on family history and life events, and on alcohol and drug use. In addition to this, *clinical and behavioural* data gathered includes genetic data (as some psychiatric disorders are held to be inherited), any diagnostic or case history, current psychiatric assessment, and performance in IQ and other cognitive tests.

The majority of this non-image data is obtained by clinicians through interviews and questionnaires, or from clinical records, with the exception of genetic data extracted from blood samples. This requires scheduling *subject visits* so that people are scanned at a convenient time to allow for a clinical interview and possibly a blood test, and also may involve contact with the subject's doctor (GP). The scanning process itself has been touched on in the brief history above, and typically involves hospital-based medical physicists who program the scanner to follow a sequence corresponding to the imaging modality (e.g. fMRI, MRI) and experimental design, and radiographers to operate the scanning process. MRI 'slice' images are commonly stored in the DICOM standard medical image file format.

The DICOM (Digital Imaging and Communications in Medicine) Standard is published by the US - based National Electrical Manufacturers Association[1]. It encompasses a file format that includes comprehensive metadata in the file header, and a TCP/IP based network communications protocol, enabling scanners to be linked to other hardware devices across a network, and integrated into a Picture Archiving and Communications System (PACS). The DICOM header is contained within the same file as the image. It comprises various metadata about that image, including the dimensions, byte order, the type of scanner or 'imaging modality' that produced it, the software version, slice thickness (number of voxels) and any compression technique used (Rorden, 2008).

[1] The DICOM Standard is available at: http://dicom.nema.org/

Image Processing

Various '*pre-processing'* steps are performed on images before they may be used in analysis (Toga 2002, Van Horn 2004). A common first step is *anonymisation* to remove metadata identifying the subject, included in the DICOM file header of each scan, before *reconstruction* of the three-dimensional brain volume from the slice data. Subsequent processing is directed at reducing the amount of 'noise' in this data. Small movements of the head are typical and image-processing software is used to correct for this. Additional *realignment* may be needed in functional studies because of the additional time factor, and differences in slice timing and other motion-related effects also need to be compensated for. Since functional scans are of relatively low resolution, they are aligned or *co-registered* with structural scans obtained at the same time. To compare images between individuals, the next step is to map the structural (or co-registered functional and structural) images to a common set of spatial coordinates for the brain. This is referred to as spatial *normalisation* or 'brain warping', and normally involves reference to a brain atlas known as the Talairach system (Toga, ibid.). The result is, for each brain volume, a set of x, y and z coordinates for each voxel[2], matched to a known neuroanatomical region with a degree of statistical significance. As a final step before statistical analysis, smoothing algorithms may be applied to improve the signal-to-noise ratio (Van Horn, 2004).

Statistical Analysis

In structural studies, statistical analysis of neuroanatomical changes can be based either on the normalised images or less commonly on the 'deformation fields' produced by that process. Analysis of the normalised brain images has traditionally involved computing the average differences in the volumes of specific areas of the brain, which has until recently depended on the identification of regions by *tracing* them- a very labour intensive manual process, partly dependent on the tracer's judgement about where a region begins and ends. Alternative approaches include *Voxel-Based Morphometry* (VBM). Instead of using mean differences in volume to compare brains, after each brain has been mapped to a common three-dimensional space it is then *segmented* into grey and white matter and cerebral fluid. This allows differences in grey or white matter density to be calculated voxel by voxel relative to a normalised brain for the group of subjects in question, and compared within and between groups (Ashburner and Friston, 2000).

In functional studies the prevalent form of analysis uses the General Linear Model to identify regression co-efficients for the predictor variables in a design matrix, which would include variables representing the demographic and clinical data of interest, as well as the task stimuli presented to the subjects when they are scanned. The statistical significance of the coefficients is calculated voxel by voxel and given as t-tests, in the form of a *statistical parametric map*. These are then used in three levels of analysis; firstly to find individual differences between subjects, and secondly the differences within the study's groups of subjects. The third level is analysis of any statistically significant differences between these groups that would test the study hypothesis, or lead to further hypotheses.

Neuroimaging analysis software typically provides *overlays* like those shown in Figure 1.3 - maps of the analysed brain volumes that use colour gradations to highlight statistically significant changes in brain activity. Researchers typically include these in

[2] The term *voxel* is used in imaging as an abbreviation of 'volumetric pixel'. Where a pixel is a unit of two-dimensional space visualised on a computer display, a voxel is a unit on a three dimensional grid used to represent a volumetric dataset or object (Kaufman et al, 2003).

publications along with tables of relevant statistics on individual or group differences, as the final results of data reduction (Van Horn 2004).

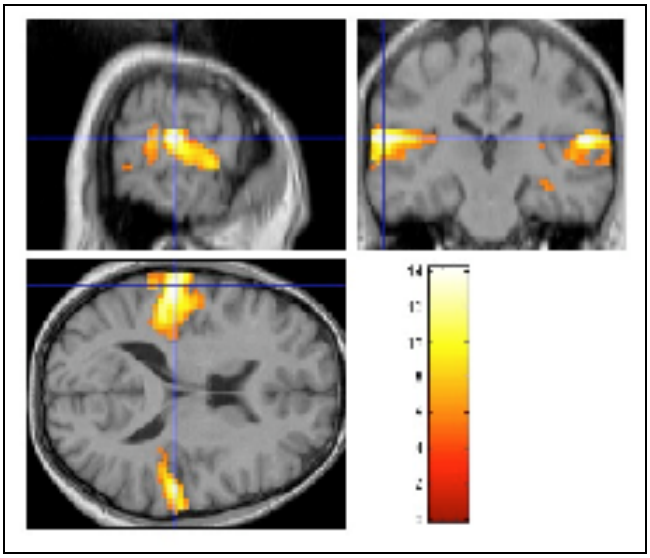


Figure 1.3 Example of overlays produced by the SPM 5 neuroimaging analysis software (Ashburner et al 2008 p.197) © Wellcome Trust Centre for Neuroimaging

1.3. Sharing and Reuse: Policy and Scientific Drivers and Constraints

Sharing neuroimaging data in psychiatry has a specific disciplinary rationale, quite apart from wider science policy considerations. The search for statistically significant correlates of neurological changes in psychiatric patients has been hampered by small sample sizes, particularly for schizophrenia (Lawrie at el 2006). This has been an important driver for 'e-infrastructure' (or in the U.S. 'cyberinfrastructure') developments aimed at integrating datasets and enabling collaboration between researchers on shared techniques (Van Horn 2004, Geddes et al 2005, Keator et al 2007). There has also been a long-running debate in the neuroimaging community on the rationale for and practicalities of data sharing. This section briefly outlines that debate and how it relates to the current policies of the major UK research funding bodies.

Benefits and Constraints of Sharing and Reuse in Neuroimaging Research

Data sharing in neuroimaging and in neuroscience generally has historically been very limited, although researchers have been more amenable to sharing analytic tools (Toga 2002, Geddes et al 2005). The arguments for sharing in neuroscience frequently draw on the experience of other disciplines; genomics and the Human Genome Project in particular are often given as exemplars of the benefits of large scale data sharing and collaborative practice; so much so that the U.S. National Institute of Health (NIH) in 1993 funded the *Human Brain Project* to underpin the development of the emerging field of *neuroinformatics* (Gardner et al, 2003)

An ongoing push for wider data sharing is linked to the growing influence of neuroinformatics in neuroscience according to Ascoli (2005). The public distribution of data and the scientific articles that refine and interpret it is held to be a catalyst for the application of informatics-related methods to neuroscience, enabling neuroinformatics to amplify the consequences and potential benefits of neuroscience research.

As in other fields the 'open science' movement has been a factor in the debate on sharing, and especially in the lead up to the NIH publishing its data sharing policy in 2003; a debate that preceded and influenced policy-making in the UK (described below). Benefits and constraints of sharing in neuroscience were characterised in a *Science* editorial by the Governing Council of the Organization for Human Brain Mapping (OHBM, 2001), and by the Human Brain Project investigators (Gardner et al, 2003) amongst others. The advantages given are-

- Comparison of findings across laboratories, and better assessment of reliability and reproducibility
- Encouraging meta-analyses that explore phenomena not apparent in individual datasets
- Access to existing data for investigators without neuroimaging facilities
- More efficient use of neuroimaging data since this is relatively expensive to collect.

Constraints that are particularly evident in the neuroscience domain are:-

- Lack of universally accepted standards for dataset structure and content, reflecting: -
- Rapid changes in methodology and knowledge, and thus the diversity of research outputs
- The highly variable demographic and clinical characteristics of the people scanned
- Lack of metadata standards for describing the conditions under which data are acquired

A key *re-use* issue according to Gardner et al (ibid.) is that it is relatively straightforward to re-use neuroscience data by performing new analyses, but also very open to misinterpretation without enough metadata to understand the original experimental context. As a result;

> "…the scope of shareable data may legitimately vary depending upon the standards and practices of different fields or techniques, and may thus include or exclude any or all of 'raw', partially processed, processed or selected datasets. Ideally shareable data should be defined as the combined experimental data and descriptive metadata needed to evaluate and/or extend the results of a study" (ibid. pp.291).

Aiming to address similar concerns about safeguarding and crediting the work of data producers, the OECD Working Group on Neuroinformatics proposed a *Legal and Policy Framework for Neuroinformatics* (OECD Working Group, 2003). This focused mainly on the intellectual property implications of sharing and re-use, to address risks of 'anti-commons effects' of sharing - the risk, that is, of publicly shared data being patented or otherwise made proprietary by re-users. To address that, the Working Group propose adopting the 'copyleft' principle of the open software movement "in order to maximize the free flow of collaborative information and allow for negotiated alternative commercialization licenses for the private sector" (ibid. pp.161)

The need to protect the confidentiality of patient data is the other major issue for these authors (OECD Working Group, OHBM, Gardner op.cit.). We come back to the legal and ethical issues surrounding this after an overview of the UK funding bodies' policies on data sharing and access.

Research Council Policy Principles

Policies on access to research data outputs are rapidly evolving at inter-governmental levels, as are those of the major UK funders of research in neuroimaging and psychiatry; the Medical Research Council and Wellcome Trust. Taking the inter-governmental developments first, the report Dealing with Data (Lyon, 1997) summarises key recent developments, such as the OECD's 2006 *Recommendation concerning access to research data from public funding*, which sets out principles and guidelines for governmental science policy and funding bodies. Since then the European Commission has begun a policy-making process on access dissemination and preservation of scientific information which may result in new government measures on research infrastructure and data access (ESFRI, 2007).

In keeping with the OECD principles, the UK Research Information Network (RIN)-sponsored by the four Higher Education funding bodies, the three National Libraries, and the seven Research Councils, has recently published principles and guidelines for stewardship of research data (RIN, 2008a), setting out five in particular: -

1. The roles and responsibilities of researchers, research institutions and funders should be defined as clearly as possible, and they should collaboratively establish a framework of codes of practice to ensure that creators and users of research data are aware of and fulfil their responsibilities in accordance with these principles.
2. Digital research data should be created and collected in accordance with applicable international standards, and the processes for selecting those to be made available to others should include proper quality assurance.
3. Digital research data should be easy to find, and access should be provided in an environment which maximises ease of use; provides credit for and protects the rights of those who have gathered or created data; and protects the rights of those who have legitimate interests in how data are made accessible and used.
4. The models and mechanisms for managing and providing access to digital research data must be both efficient and cost-effective in the use of public and other funds.
5. Digital research data of long term value arising from current and future research should be preserved and remain accessible for current and future generations.

The MRC and Wellcome Trust have published data sharing and access policies advocating similar principles and informed by several key reports drawing on the US and international policy debates; (Lowrance 2006), and the Joint Data Standards Study Report (DAC/BRC/NeSC, 2005). The MRC policy statements (MRC, 2007) include a *Data Sharing and Preservation Policy* and a more detailed *Data Access Policy*. The Wellcome Trust covers similar ground in its *Policy on Data Management and Sharing* (Wellcome Trust, 2007). These statements set out rights and obligations of data producers or custodians and include:-

- A requirement to make datasets available with few restrictions, while balancing this with the need to protect the personal data of research subjects.
- A requirement to produce a plan for sharing and managing data as a condition for funding.
- Provision for a period of exclusive access by the data producer, and for measures to protect intellectual property rights.

The policies recognise that there is no current provision of a data centre infrastructure to support community-wide archiving and curation (although Wellcome Trust point to various major databases they fund). Responsibility for archiving is largely the concern of data producers and custodians together with their institutions, and the MRC identify a retention period of 10 years from the end of a funded project. The Wellcome Trust's policy explicitly identifies that funding for data management is a legitimate component of a research grant, and that costs of equipment and databases may be approved.

Ethical Concerns and Privacy Legislation

The MRC and Wellcome Trust policies have a strong focus on compliance with ethical, statutory and other regulatory requirements that arise for researchers using data or tissue from human participants. These are briefly outlined here (see also MRC, 2007), with background on some ethical issues specific to neuroimaging.

The MRC provides researchers with guidance and toolkits on the requirements of Data Protection legislation and the Human Tissue Acts (MRC, 2007). As well as these, researchers and their institutions are required to comply with the NHS research ethics approval process. While the arrangements vary across different parts of the UK, each places stringent limits at the outset of a research project on the kinds of data that may be gathered, how that data may be processed and what may be retained. An important factor in studies including psychiatric patients is that any risk of distressing them must be minimised.

Researchers are obliged to anonymise any data that may be identified with a research subject before it can be shared with other researchers. A key issue here is whether informed consent to share the data with other researchers has been obtained from the data subject. Where it has not the approval process is complicated and the publication *Personal Information in Medical Research* (MRC, 2000) provides guidance on this. In addition any potential re-user of data that is not anonymised must seek research ethics committee approval. This also applies even within a research group; for example a research student seeking to analyse existing MRI images for a dissertation would be exempt only if the images were anonymous to the research team that holds them, and any coded link to personal data held by a larger research database or register.

While the above apply to any medical research, neuroimaging has been the subject of enough legal and ethical debate to attract its own label; '*neuroethics*' (e.g. Racine and Illes 2006, Fukushi et al, 2007, Goldberg 2007, Khoshbin and Khoshbin 2007, Tancredi and Brodie 2007). According to Kulynych (2002) this is a sign of "…the capacity of imaging technology to animate mind–brain relationships in ways that compel us to re-examine concepts of identity, personal responsibility, and criminal culpability." (pp. 345). Ensuing neuroethical considerations for researchers and ethics committees include: -

- Potential health risks to subjects from MRI scanning itself, i.e. distress and discomfort from confinement and noise, plus possibilities of physical injury.
- The possibility that an imaging study might predict psychological illness, cognitive impairment or reveal some other serious medical condition, with the implications for possible treatments (or lack of those), and the subject's distress and insurability.
- Given the large costs of imaging studies and equipment, the possible conflicts of interest for research sponsored by manufacturers and drug companies. (Kulynych, 2002, 2007).

There is also significant concern around the identifiability of MRI images. The reconstruction of three-dimensional surface views of the brain from structural images

produces images with recognisable faces. This means that even when all other identifying information is removed from the scan header and any accompanying data files, the MRI scan is still potentially identifiable, and might be matched against a database of photographs or digital images of known individuals, including by automatic facial identification techniques (Kulynych, 2002).

Even with the development of techniques for stripping facial features, the possibilities for identification do not stop there; the OECD Neuroinformatics Working Group reported in 2002 that "it is likely that … neuroanatomy could serve as a unique "fingerprint" for identifying individuals." (OECD Working Group, 2003, pp.160).

1.4 Neuroimaging Repository and Infrastructure Developments

Data and Publication Repositories

The risks and rewards of public neuroimaging databases are substantial considering the high potential for re-using neuroimaging data, coupled with the high ethical and regulatory needs to safeguard the confidentiality of potentially very vulnerable people. These factors contribute to a landscape that has some large and ambitious public database projects, but very few established repositories or archives, and a history of low investment in database technologies at the level of individual laboratories.

Considering the individual laboratory first, according to Toga (2002) relatively simple archival catalogues of images are often used in laboratories. However Bug and Nissanov (2003) say that labs rarely make use of "even the most rudimentary tools available" to manage image data, and Geddes et al (2006) note that imaging data is typically stored in file directories. They also state "data curation in neuroimaging research tends to be poor" (p. 360).

While the UK lacks established data centres from which individual laboratories might receive support for data archiving or shared repository services the MRC has funded several e-Science projects to develop innovative services in this area (see below). It is also (at time of writing) establishing a data support service, and currently publishes a *Cohort Dataset Directory For Mental Health Researchers*, in conjunction with the Mental Health Research Network[3]

Access to research outputs is facilitated through the MRC & Wellcome Trust 'open access' policies, which require electronic copies of any research papers from work they have funded and have been accepted for publication in a peer-reviewed journal, to be made freely available from PubMed Central (PMC) and other PMC International (PMCI) repositories such as UKPMC. All deposited papers must be made freely accessible from the PMC and other PMCI repositories as soon as possible, and in any event within six months of the journal publisher's official date of final publication.

Initiatives to support large-scale publication of *datasets* have been embarked on by various U.S. and international organisations. They vary in the purpose and organisation of the content. Some databases are provided as canonical reference data in the form of web-based brain atlases and coordinate systems, used by researchers to align their results with templates and statistics representing the norms of brain structure or function. Other databases are intended to provide the primary data or derived results from specific studies (Toga, op.cit.)

[3] MRC & MHRN Cohort Dataset Directory For Mental Health Researchers is available at: http://developers.psygrid.org:9080/mrc-dsc-web/app/directoryHome

Brain atlases include the *Whole Brain Atlas* [4](Harvard University), which provides templates combining imaging and clinical reference data. By contrast the European Commission funded the *NeuroGenerator* project[5] to provide derived databases of statistical parametric images in a form suitable for meta-analysis.

Meta-analysis capabilities are increasingly a goal of several longer-established initiatives funded by the U.S. National Institute of Health, to provide repositories of neuroimaging experiments and metadata. These are: -

- BrainMap [6]; an online database of functional experiments with metadata on the brain coordinates of the 'activation locations' reported in publications, and tools to perform meta-analyses on those locations.
- The fMRI Data Center (fMRIDC) [7]; a public repository of fMRI studies that links peer-reviewed papers published in the *Journal of Cognitive Studies*, with experimental metadata and de-identified image data.
- FBIRN Human Imaging Database (HID) [8]; the main goal of fBIRN (Functional Bioinformatics Research Network) is to develop tools to facilitate multi-site functional MRI studies. The HID is part of the BIRN Data Repository and provides access to image data, statistical results and tools to analyse these.

International organisations promoting database development in neuroimaging include: -

- The Society for Neuroscience; through a Neuroscience Database Gateway [9], provides a curated catalogue of databases of experimental data and research materials.
- INCF (International Neuroinformatics Coordinating Facility) [10], initiated as a result of the OECD Neuroinformatics Working Group, aims to coordinate and foster activities through development and maintenance of database and computational infrastructure and support mechanisms for neuroscience applications.

Standards and other Infrastructure Developments

Efforts to promote meta-analysis are currently made difficult by the lack of standardisation in experimental methods and description mentioned earlier (Forsberg and Roland, 2007). These have therefore been a major focus of neuroimaging e-science or (in US terminology) cyber-infrastructure initiatives. In the UK e-Science programme the MRC-funded Neurogrid, PsyGrid and NeuroPsygrid projects exemplify work on technology infrastructure for neuroimaging psychiatric research. Others are also outlined in this section; the Neurobase project in France, and the US - based BIRN initiatives.

In the US the term 'cyberinfrastructure' has been broadly defined, according to the Atkins Report, to refer to "layers that sit between base technology (a computer science concern) and discipline-specific science" (Edwards et al, 2007). In the UK, infrastructure for e-science, or 'e-infrastructure' has had a more specific focus on "the distributed computing infrastructure that provides shared access to large data collections, advanced ICT tools for data analysis, large-scale computing resources and high performance visualisation. It embraces networks, grids, data centres and collaborative environments". (OSI, 2007).

[4] Whole Brain Atlas is available at: http://www.med.harvard.edu/AANLIB/home.html
[5] Neurogenerator is available at: http://www.neurogenerator.org/
[6] BrainMap is available at: http://brainmap.org/
[7] fMRIDC is available at: http://www.fmridc.org
[8] fBIRN Human Imaging Database is available at: http://fbirnbdr.nbirn.net:8080/BDR/
[9] Neuroscience Database Gateway is available at: http://www.sfn.org/
[10] INCF is available at: http://www.incf.org/

Neuroimaging e-science projects have focused on enabling shared access to large data collections, by integrating datasets at various points in their lifecycle. In the psychiatric neuroimaging field many e-science projects have a particular focus on *schizophrenia* research and better enabling it to predict the onset of psychotic symptoms in those most vulnerable to the disorder.

Research on schizophrenia embodies the quest for larger samples and many of the constraints on multi-site studies described earlier. Publications from *Neurogrid* (e.g. Ure et al 2007) describe the population differences such as ethnic differences in brain shape, and site differences in data collection, coding and collation that confound analysis of pooled data. The Neurogrid project includes a 'psychosis exemplar'. It aims to create a Grid-based infrastructure to connect neuro-imaging centres and develop Grid-based data analysis tools and services that aim to improve diagnostic performance, to enable differences between images from different scanners to be compensated for and to allow quality and consistency verification (Geddes et al, 2005).

The *Psygrid* project has a similar focus on psychosis and on enabling large enough cohort studies to reliably address research questions. It has a more explicit focus on clinical practice and on providing infrastructure to support 'early intervention' in patient treatment (Ainsworth et al 2007). The *NeuroPsygrid* project builds on the overlapping ground of this and Neurogrid, to combine datasets and extend work on a shared metadata model. This work aims to develop an ontology of terms for psychosis to address the disparate scales that researchers have used to describe symptoms, and to make this work consistent with work on ontologies in the BIRN project, which already provides an ontology (BIRNLex) for integration of experimental data from studies (Kola et al, 2008). Ontology development has also been aim of European projects such as the French Government sponsored Neurobase (Barillot at el, 2005).

Development of *metadata schema* has been a major element of work in BIRN. This includes XCEDE an XML-based data exchange schema representing metadata of various types. These include descriptions of experiments and subject visits, time-series data from fMRI studies, and provenance i.e. documentation of the analysis workflow steps taken. (Keator et al, 2007).

Human and technology infrastructures for data access

Significant characteristics of neuroimaging e-infrastructure also include its less technical ones. Lee et al (2006) attribute the view that '*human infrastructure'* is the most critical aspect of cyberinfrastructure to Fran Berman, director of the US National Partnership for Advanced Computational Infrastructure and the San Diego Supercomputing Center. Their case study of the fBIRN network questions the idea that 'distributed teamwork' is a sufficient basis for understanding how large-scale collaboration is accomplished. They argue instead that the more fluid form of working they found in fBIRN can better be understood by looking at how such collaborations 'blend' such local concerns as institutional prestige, organizational relationships, access to appropriate scientific data, and subjects.

Differing *access models* are characteristic of neuroimaging infrastructure projects. For example BIRN's Human Imaging Database allows public access with minimal checks on whether the database user is a bona-fide researcher, whereas the PsyGrid project provides limited access using a role-based access model (Ainsworth et al, 2007).

Factors influencing data access levels are identified in Lowrance's report on medical data access (Lowrance, 2006, pp. 12-14). This points out that "open access" may refer

to web-based access that is without any restriction, or to data that is open to *applications* for access. Similarly, there are degrees of collaboration and these may change over time; ranging from close colleagues, other members of the same institution, external researchers known to and respected by the custodians, or others who have no prior relationship at all with the custodians. Access policies for determining which applications are legitimate may involve a wide range of terms and conditions concerning: -

1) Confirmation of professional competence
2) Screening of the scientific merit and relevance of proposed collaborations
3) Specification of what is to be provided
4) The terms of consent used when collecting data from subjects
5) Purpose limitations such as educational or research use
6) Confidentiality and statutory limitations
7) Research ethics approval
8) Linking to other data the applicant holds
9) Re-contacting the data subjects
10) Maintaining or enriching the quality of the resource.
11) Publication or co-publication requirements
12) Archiving requirements
13) Assigning or waiving of intellectual property (IP) rights
14) Responding if consent is withdrawn by a research subject
15) Prioritisation of access to limited resources, e.g. by committee decision
16) Access fees or royalties
17) Returning or destroying materials
18) Trans-border enforcement
19) Termination clause covering cessation or transfer of primary custodianship
20) Standard legal disclaimers of responsibility for errors, inaccuracies, or for consequences of use.

One of the challenges for neuroimaging e-infrastructure projects is to address the range

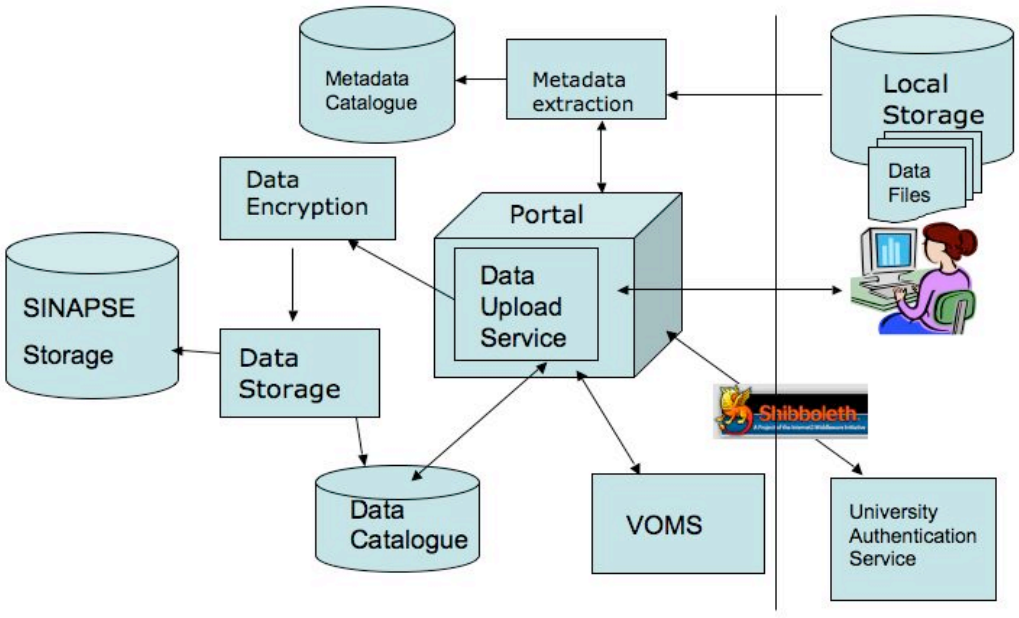


Figure 1.4 Architecture for the SINAPSE Project (source: Rodriguez et al 2008)

of constraints on access indicated above; in fact Lowrance (ibid.,p.20) identifies confidentiality and anonymisation as one of the 'issue clusters' most in need of attention for data sharing in medical research. Meanwhile, various data storage and sharing models have been deployed in the service of neuroimaging and access to its related data.

While in BIRN and Neurogrid datasets are held in federated grid storage, centralised storage architecture is used in PsyGrid (Ainsworth et al 2007) and in the SINAPSE Project (Rodriguez et al 2008). The latter has broad aims to improve the infrastructure for neuroimaging research in Scotland. Figure 1.4 illustrates the proposed SINAPSE architecture. This seeks to automate data anonymisation, provide effective trade-offs between security and usability, and a modular framework that is deployed using centralised storage and computation resources in order to reduce the overheads these would present to the collaborating labs.

Collaboration, trust and the organisational and technical infrastructures to support it were identified in this chapter as critical ones for digital curation- and for data integration in neuroimaging, and they were a recurring theme in the case study that followed.

2. Case Study Methodology

The term ‘case study’ is widely used and embraces various methods, the common theme being a focus on describing, explaining or changing ‘real world’ activities working with those involved using some combination of participant observation and other social science methods. The approach outlined here is consistent with the widely accepted definition by Yin of the case study as an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident (Yin 2002).

The approach aimed to develop field research practice in digital curation along the well-established lines of CSCW (Computer Supported Collaborative Work) which similarly integrates the ‘systemic analysis’ of current practices with ‘appreciative intervention’ to introduce change. Helena Karasti, one of the few in the CSCW community to have applied such as approach to data curation describes the main challenge as bridging “discrepancies between ethnographic studies of work and system design, such as descriptive vs. prescriptive, particular vs. general, concrete vs. abstract, present vs. future, and understanding vs. intervention.” (Karasti, 2001 p.239). This form of ‘sociotechnical’ research seeks to bridge academic sociological studies of organisation with the pragmatic aims of informatics disciplines to ‘improve’ organisation.

More specifically, as one of the concerns of digital curation is preservation of ‘context’ and ‘provenance information’, a principled basis was sought for understanding and documenting those aspects of practice that a research community deems significant enough to record or would otherwise miss, at the same time as offering them pragmatic help to address questions about how, where and when to record them.

The study therefore combined ethnographic and action research methods to (respectively) describe current practices in the host research group, and to identify preservation and curation issues and support the group to further develop these capabilities. This contrasts with quantitative survey-based social science approaches of testing a-priori hypotheses by predefining the salient variables at the outset, measuring correlations between these and generalising statistically from a sample of participants to the population of interest.

An overview of the approach follows with some reflections on its limitations and improvements that might be made.

2.1 Action Research

Action research stresses intervention by researchers in the problem domain, to collaboratively learn how to address it with those affected by and responsible for change. The essential characteristics (Davison et al 2004) are a cyclic process of identifying problems or situations in need of improvement; agreeing relevant courses of action, intervening to take the appropriate action, and reflecting on its outcomes. Action research is not driven by theory at the outset (e.g. to formulate and test hypotheses) although it may be guided and informed by theory at the data collection and analysis stage (ibid). The approach has many variations, but SSM - Soft Systems Methodology (Checkland and Scholes, 1999) is one that is commonly applied to information systems contexts. SSM is intended for use in any problematic situation that involves ‘would-be problem-solvers’ identifying feasible and desirable action.

The methodology involves identifying and modelling ‘relevant systems of human activity’ (i.e. relevant to addressing the problematic situation), and stresses attention to issues

that arise from intervening in the situation. Modelling 'relevant systems' involves first identifying a 'root definition' comprising a 'transformation' of inputs into outputs, the 'actors' who would perform it, 'owners' who may stop it, 'customers' or stakeholders, i.e. those benefiting or otherwise affected, and the worldviews underpinning the proposed change and environmental constraints on that change (ibid.).

Simple conceptual models are then drawn to depict a sequence of activities relevant to accomplishing change, and the criteria actors would use to assess those activities. The purpose of this very loosely structured form of modelling is to reflect on alternative ways of framing intervention and identify a feasible and desirable course of action (ibid.). In the Neuroimaging case this helped plan the two main phases of the study. The approach was used more explicitly towards the end, to draft the 'roles and goals' of a system to support documentation of the Neuroimaging Group's studies.

2.1.2 Risk Assessment

Action research in this study focused more specifically on risk assessment, to involve Neuroimaging group members in identifying and scoping 'challenges' to data preservation and curation and what might be done about these. The DRAMBORA toolkit - *Digital repository audit method based on risk assessment* (DCC/DPE, 2007) is designed as a toolkit for *repository administrators* to identify and manage risks to datasets they hold. Here it was adapted to assess the data preservation risks of a relatively small research group, in the spirit of the foreword to the toolkit, which points out that: -

> "…most current digital repositories, and most databases and collections used to help curate scientific data, do not have specific mandates for long term preservation, nor do they have the necessary long-term budgets. Instead, they are mandated to support access and re-use in the near-term future. Long-term preservation may be one of their aims, or at least hopes and wishes, but it is not (yet) a responsibility. *Much of the work on attributes and criteria … is not oriented to this large group of repositories, although parts of it may prove helpful*." (ibid. emphasis added).

The rationale for this approach was the availability of the DRAMBORA 'self-assessment' toolkit, and its apparent suitability given that, within Neuroimaging Group, risk assessment was seen as offering potential to involve clinicians and other members in something they were collectively responsible for, rather than an exercise in identifying changes to technical or administrative procedures. In practice around half of the group were interviewed about 'data curation challenges' and only those few who actively involved in group-wide curation were approached to take part in assessing the risks identified.

DRAMBORA also proved difficult to apply in a research group context, where the data potentially at-risk was being actively acquired and used for its creators' needs. The methodology is tied to formally defined repository objectives, functions and roles, rather than to research groups who do not currently have these but want to develop their data archive function. For example in the DRAMBORA approach risk *consequences* are characterised in terms of risks to *data*. It was almost counter-intuitive to translate risks to the research objectives of the data creator (Neuroimaging Group) into risks to its data, whereas in an archival organisation these objectives might be more straightforwardly linked.

Other issues were that:-

1. The risk probability categories make it difficult to deal with an event that is likely every 3 years, i.e. midway between 'once every year' and 'once every 5 years', even

although research departments often have to account for spending needed to mitigate risks in 3 year periods.

2. Risk impacts are categorised in terms of the extent of data loss, rather than any loss of understandability, usability or value of the datasets or other assets.

It seemed more appropriate to envisage risks to data in an active research group in terms of -

- Impacts on data *usability* that range from those isolated to a dataset (corresponding to a project on someone's personal hard disk, or in a server directory) to those that are widespread (corresponding to everything held on a server).
- Impacts on *value* that range from relatively low value datasets to those that are irreplaceable, where 'value' is some combination of the acquisition cost; the potential for the primary user to get highly rated research outputs from the dataset; and the potential for others in the research community to do that (ie re-use).

The probability & impact ratings that were used were: -

Probability

5	Very high probability, at least once per month
4	High probability, at least once per year
3	Medium probability, every 1 to 2 years
2	Low probability, every 3-5 years
1	Very low probability, every 6- 10 years

Risk Impact

1	Superficial - Isolated loss of dataset usability or value, which can be recovered
2	Medium - Isolated loss of dataset usability or value, some unrecoverable
3	High - Widespread loss of dataset usability or value, some unrecoverable.
4	Considerable - Widespread loss of dataset usability or value, mostly unrecoverable.
5	Cataclysmic - Total and unrecoverable loss of datasets usability or value

DRAMBORA could be more comprehensively adapted to research groups, adopting aspects of the DCC Curation Lifecycle (which was developed around the same time as this case study) and the Data Audit Framework methodology, which was published shortly after it. These would respectively offer a more relevant basis for identifying curation activities and digital data assets in the earlier stages of risk analysis.

2.3 Ethnographic Fieldwork

Ethnographic fieldwork methods have complemented action research in the CSCW (Computer Supported Collaborative Work) field, by describing in their context the working practices that would be affected by envisaged systems development work. Ethnography refers more to the aim of 'richly describing' practice, typically using semi-structured interviews and participant observation, rather than to a single analytic approach. Two common approaches in systems contexts are 'ethnomethodologically-informed' ethnography, and 'grounded theory' (see e.g. Randall et al 2007 for comparison of the above). The relevance of these approaches to the current case is discussed below: -

Describing 'data practices'

Ethnographic analysis that is informed by ethnomethodology (the study of 'people's methods') highlights through close reading of field notes or transcripts how practitioners use their shared knowledge to collaboratively accomplish tasks and account for their actions (ibid). SCARP is concerned with describing practices of research data sharing, curation, archiving, re-use and preservation. These are more likely to be embedded in the day-to-day working of a research group than identified with distinct curation roles that one would expect to find in a digital library or archival organisation. It was therefore important to identify *how* data sharing (et al) practices are a feature of group members' interactions; that is with how those interactions demonstrate and accomplish care for data, how in their shared efforts to develop professional knowledge they make data accessible to others, and how past work is re- purposed. Material on this was gathered from interactions the SCARP researcher was party to, i.e. observations recorded in field notes, and from interviews.

This form of analysis provided an understanding of how the 'stewardship' of data is practically accomplished and how and when researchers document the context of their work for future reference. Much of the material for analysis came from observation of weekly group meetings over 12 weeks and from the semi-structured interviews described below.

The sufficiency of the analysis as 'ethnomethodologically-informed ethnography' deserves comment. Although a detailed discussion of ethnomethodology is beyond the scope of this report two relevant aspects of it are the notion of 'unique adequacy' and the role of theory. It is important to note that ethnomethodology aims for understanding and description of 'members' practices' - where 'member' denotes any 'ordinary member of society'. According to the principle of *unique adequacy* (Garfinkel 2002, pp. 175-6), to understand and describe a practice it is necessary for the researcher (in ethnomethodology) to be able to perform it competently.

The unique adequacy principle can be regarded as an aspiration rather than a categorical requirement for ethnomethodological studies (Lynch and Cole, 2005) implying that it is sufficient to acquire a level of 'interactional expertise' (Collins and Evans, 2002), i.e. an ability to converse intelligibly with practitioners and describe the rudiments of their technical practices, while falling short of an ability to do them (op.cit.). The current study only touched on this level of expertise; as the main report mentions, it was difficult to follow and meaningfully record the substance of the discussions between neuroimaging researchers in the weekly meetings that were observed. It was this difficulty that steered the study into the area of how junior researchers' get round their own difficulties picking up the inter-disciplinary terminology used by their more expert colleagues, and the role of the meetings in providing for that. However the analysis of how this flowed in meetings was limited by having only a superficial comprehension of the hosts' professional language, as well as by the relatively limited period of fieldwork.

While some ethnomethodologically-informed ethnographers aspire to competent performance of whatever practices are being studied, explanatory theory is not an aim of this approach, which prefers to describe the use made of members' own categories rather than use sociological concepts in place of these. Theory is often regarded by ethnomethodologists as an obstacle to rigorous sociological description, since sociologists tend to use theory *instead of* describing the actions and interactions that make a social setting or way of working recognisable to its members, yielding sociological explanations less insightful than any those members could come up with themselves (e.g. Crabtree et al, 2000, White et al 2006).

It is worth reflecting on the use of Weick and Robert's (1993) term 'heedful interaction' in the current case. The main report identifies various ways that Neuroimaging Group members' interacted in ways that accommodated differences between their research backgrounds and levels of expertise. In doing so it does not seek to explain those interactions in terms of Weick and Roberts' theory of 'collective mind', to which they link the term heedful interaction. In this study the notion of 'heedfulness' is used in its ordinary language sense, and as a pointer to other aspects of the Group's ways of working rather than as an abstract model that would stand in place of them. Admittedly the Neuroimaging Group's 'heedful interaction' *is* used here as an explanation for their (reported) low levels of data loss, and Weick and Roberts' notion that 'collective mind' explains organisation reliability was in a sense 'applied to' the situation to open up a line of enquiry. However the report does not offer 'collective mind' as a theoretical account of that situation – only using 'heedful interaction' as a shorthand notation for explanations that members gave themselves – that weekly meetings or informal approaches were, for practical purposes, normally satisfactory ways to learn about a dataset created by a colleague – and that improvements to these ordinary methods were largely needed on the rationale of departing and retiring or unavailable colleagues, and the demands of research funders.

Thematic analysis

SCARP is concerned with analysing how data practices are locally organised in the research groups studied, but it is also concerned with how these relate to broader practical and academic concerns. There was therefore a need to identify common themes from interview data, for example to relate current activities to the DCC curation lifecycle model.

The approach was superficially similar to 'grounded theory', in that themes identified in initial interviews were related to literature (on data anonymisation for example), and those reflecting the interviewees' views and concerns were followed through in the risk assessment and 'looked for' in observations of meetings. Initial answers to questions were compared and summarised, and these themes were elaborated in further discussion and then related to those identified in field notes of meetings. Following this, a questionnaire was used to gauge the scale of some of the risks perceived among the group, their preferences for documenting data, and their attitudes to data sharing, which were explored in further interviews. The resemblance to grounded theory ends there however- it was not the intention to develop a coding scheme or theory from the analysis.

Semi- structured interviews were carried out between November 2007 and March 2008. A total of 20 interviews were carried out with Professor of Neuroimaging Stephen Lawrie and Group members whose roles included senior lecturer, lecturer, post-doctoral research assistant, research assistant, PhD student, and systems manager. The interviews were initially based on a topic guide included in Appendix 2. This was compiled at the outset of the study, drawing on relevant previous studies. The topic guide is relatively highly structured, but the interviews themselves were not; questions were selected according to the roles, interests and experience of each interviewee so that in the later 'immersive' phase of the study they were less formal and reflected issues arising from earlier interviews and group meetings.

Theory and case study method

There has only been limited scope here to consider the place of social/organisational theories in case studies of research practice, and address critical questions such as the methods of engagement appropriate to the varying levels of familiarity with researchers' working practices that can feasibly be attained in several months of field work. Future

reports should consider these, to further develop methodologies for understanding practices and identifying requirements at the 'coal-face' of research data curation.

3. References

Ainsworth, J., Harper, R., Bridges,L., Whelan, P. Vance, W. and Buchan, I. (2007) 'The Challenges of Clinical e-Science: Lessons Learned from PsyGrid' *Proceedings e-Science All Hands Meeting 2007*.

Ascoli, G. (2005) 'Looking Forward to Open Access' *Neuroinformatics*, 3(1), pp.1-3.

Ashburner, J. and Friston, K. (2000) 'Voxel-Based Morphometry-The Methods' *NeuroImage* 11, pp. 805– 821

Barillot, C., Benali, H. Dameron, O., Dojat, M. , Gaignard, A. Gibaud,B., Kinkingnéhun, S., Matsumoto, J., Pélégrini-Issac, M., Simon, E., Temal, L. and Valabregue, R. (2005) *Federating Distributed and Heterogeneous Information Sources in Neuroimaging: The NeuroBase Project* INRIA Research Report N° 1712. Available at: www.irisa.fr/visages/neurobase

Bug, W. and Nissanov, J. (2003) 'A Guide to Building Image-Centric Databases' *Neuroinformatics* ,1, pp. 359–378

CASPAR Project (2006) D4101 User Requirements and Scenario Specifications available at:

Cockburn, A. (2001) *Writing Effective Use Cases* London: Addison Wesley

Collins, H. and Evans, R. (2002) 'The Third Wave of Science Studies: Studies of Expertise and Experience', *Social Studies of Science* 32(2), pp. 235-96.

Corti, L. & Wright, M. (2002) *MRC Population Data Archiving and Access Project: Consultants' Report*, Chichester, UKDA available at: www.mrc.ac.uk/PolicyGuidance/EthicsAndGovernance/DataSharing/DataSharingReports/index.htm

Crabtree, A., O'Brien, J., Nichols, D., Rouncefield, M., and Twidale, M. (2000) "Ethnomethodologically informed ethnography and information systems design", *JASIST*, vol. 51(7), pp. 666-682.

DAC/BRC/NeSC (2005) *Large-scale data sharing in the life sciences: the Joint Data Standards Study Repor*t available at: www.mrc.ac.uk/PolicyGuidance/EthicsAndGovernance/DataSharing/DataSharingReports/index.htm

DCC (2008) *Sharing Medical Data- The Legal Considerations* available at: http://www.dcc.ac.uk/resource/legal-watch/

Day, M. (2007) Curating our Digital Scientific Heritage: a Global Collaborative Challenge, *paper presented at the 3rd International Digital Curation Conference*, Washington, D.C., December 11-13, 2007

Digital Curation Centre, Digital Preservation Europe (2007): Digital repository audit method based on risk assessment (DRAMBORA), version 1.0, http://www.repositoryaudit.eu/

Edwards, P., Jackson, S., Bowker, G. and Knobel, C. (2007) 'Understanding Infrastructure: Dynamics, Tensions, and Design' National Science Foundation, available at: http://hdl.handle.net/2027.42/49353

Ellard,D. (2007) 'The History of Magnetic Resonance Imaging' available at: http://www.isbe.man.ac.uk/personal/dellard/dje/history_mri/history%20of%20mri.htm

Ellaway, R., Cameron, H. and Ross, M. (2006) *Clinical Recordings for Academic Non-clinical Settings: CHERRI Project Report* available at: www.jisc.ac.uk/uploaded_documents/cherri-report_final.pdf

Forsberg, L. and Roland, P. (2007) *Proceedings of 1st INCF Workshop on NeuroImaging Database Integration*. August 30-31, 2007 International Neuroinformatics Coordinating Facility Secretariat; Stockholm, Sweden.

Fukushi, T., Sakura, O. and Koizumi, H. (2007) 'Ethical considerations of neuroscience research: The perspectives on neuroethics in Japan' *Neuroscience Research* 57 (2007) pp. 10–16.

Garfinkel, H. (2002) *Ethnomethodology's Program: Working out Durkheim's Aphorism*, Rowman & Littlefield: Lanham, MD (U.S.)

Geddes, J., Mackay, C., Lloyd, S., Simpson, A., Power, D., Russell, D., Katzarova, M., Rossor, M., Fox, N., Fletcher, J., Hill, D., McLeish, K., Hajnal, J. V., Lawrie, S., Job, D., McIntosh, A., Wardlaw, J., Sandercock, P., Palmer, J., Perry, D., Procter, R., Ure, J., Bath, P., and Watson, G. (2006). 'The Challenges of Developing a Collaborative Data and Compute Grid for Neurosciences'. In *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (June 22 - 23, 2006)*. CBMS. IEEE Computer Society, Washington, DC, 81-86. DOI= http://dx.doi.org/10.1109/CBMS.2006.156

Goldberg, S. (2007) 'MRIs and the Perception of Risk' *American Journal of Law & Medicine* (33)3 pp. 229-237

Gorman, M. (2002) 'Levels of Expertise and Trading Zones: A Framework for Multidisciplinary Collaboration' Social Studies of Science, Vol. 32, No. 5/6. pp. 933-938.

Hartswood, M. Procter, R. Rouncefield, M., Slack, R., Soutter, J. and Voss, A. (2003) 'Repairing' the Machine: A Case Study of the Evaluation of Computer-Aided Detection Tools in Breast Screening . *ECSCW 2003: Proceedings of the Eighth European Conference on Computer Supported Cooperative Work* , September 2003. pp. 375–394

Hilder, W. (2005) Medical Research Council (MRC) 'IT Data Storage/Preservation Group Remit' available at: http://www.dcc.ac.uk/events/mdb-2005/hilder.php

Job, D., Whalley, H., McIntosh, A., Owens, D., Johnstone, E., and Lawrie, S. (2006) 'Grey matter changes can improve the prediction of schizophrenia in subjects at high risk' *BMC Medicine* 2006, 4(29) doi:10.1186/1741-7015-4-29

Johnson, K and Becker, A. (1999). 'The Whole Brain Atlas: Neuroimaging Primer ' available at: http://www.med.harvard.edu/AANLIB/home.html

Karasti, H. (2001) 'Bridging Work Practice and System Design' *Computer Supported Cooperative Work* (10) pp. 211–246.

Kaufman, A, Cohen, D, and Yagel, R. (1993) 'Volume Graphics' IEEE Computer, 26(7) July 1993, pp. 51-64.

Keator D, Gadde S, Grethe J, Taylor D, Potkin S, FIRST BIRN. (2006) 'A general XML schema and SPM toolbox for storage of neuro-imaging results and anatomical labels' *Neuroinformatics*. 4(2), pp.199-212.

Khoshbin, L. and Khoshbin, S. (2007) 'Imaging the Mind, Minding the Image: An Historical Introduction to Brain Imaging and the Law' *American Journal of Law & Medicine* (33)3 pp. 171-192

Kulynych, J. (2002) 'Legal and ethical issues in neuroimaging research: human subjects protection, medical privacy, and the public communication of research results' *Brain and Cognition* 50, pp. 345–357

Kulynych, J. (2007) 'The Regulation of MR Neuroimaging Research: Disentangling the Gordian Knot' *American Journal of Law & Medicine* (33)3 pp.295-317

Lawrie, S. Weinberger, D., and Johnstone, E. (2005) *Schizophrenia: From Neuroimaging to Neuroscience* Oxford: Oxford University Press.

Lee, C., Dourish, P. and Mark, G. (2006) 'The Human Infrastructure of Cyberinfrastructure' *Proceedings CSCW'06* New York: ACM

Lord, P. and Macdonald, A. (2003) A. *e-Science Curation Report* available at: www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

Lowrance, W. (2006) *Access to Collections of Data and Materials for Health Research; A report to the Medical Research Council and the Wellcome Trust* available at: www.wellcome.ac.uk/accessreport and www.mrc.ac.uk/research_collection_access

Lymer, K., Job, D., Moorhead, W., McInstosh, A., Owens, D., Johnstone, E., and Lawrie, S. (2006) 'Brain–behaviour relationships in people at high genetic risk of schizophrenia' *NeuroImage* 33 pp. 275 – 285

Lynch, M. and Cole, S. 'Science and Technology Studies on Trial' (2005) *Social Studies of Science*, 35(2), pp.269-311

MRC (2000) *Personal Information in Medical Research* available at: http://www.mrc.ac.uk/PolicyGuidance/index.htm

MRC (2007) *Policy and Guidance* available at: http://www.mrc.ac.uk/PolicyGuidance/index.htm

McChesney, S. Gallagher, S. (2004) 'Communication and co-ordination practices in software engineering projects' *Information and Software Technology* (46) 473–489

Miyapuram, K., Pammi, C., Ahmed, A., and Bapi, S. (2007) Neuroinformatics Tools for Functional MRI: Experimental Design and Data Analysis (unpublished pre-print) Cogprints: available at http://cogprints.org/5485/ (May 2008)

OHBM (2001) 'Neuroimaging Databases' *Science* 1 June 2001: Vol. 292. no. 5522, pp. 1673 – 1676.

OSI (2007) 'OSI e-Infrastructure Working Group, Developing the UK's e-infrastructure for science and innovation Office of Science and Innovation'. available at: http://www.nesc.ac.uk/documents/OSI/index.html

Patel, M. (2008) DCC Data Centres Synthesis Study (forthcoming)

Pekar, K. (2006) 'A Brief Introduction to Functional MRI' *IEEE Engineering in Biology and Medicine* March/April 2006, pp. 24-26

Racine, E. and Illes, J. (2006) 'Neuroethical Responsibilities' *Canadian Journal of Neurological Sciences* 33(3), 2006, pp. 269-277

Randall, D., Harper, R. and Rouncefield, M (2007) *Fieldwork for Design: Theory and Practice* London: Springer-Verlag

Ribes, D. and Finholt, T. A. (2007): 'Planning infrastructure for the long-term: Learning from cases in the natural sciences' in Proceedings of the Third International Conference on e-Social Science October 7-9, 2007, Ann Arbor, Michigan, USA.

Rodriguez, D. Carpenter, T. van Hemert, J and Wardlaw, J. (2008) 'E-Infrastructure for Data Sharing in the SINAPSE Project' Proceedings e-Science All Hands Meeting 2008, available at: http://www.allhands.org.uk/programme/index.html

Rorden, C. (2008) Introduction to DICOM medical image format. Available at: http://www.sph.sc.edu/comd/rorden/dicom.html
Star, S.L. and Ruhleder, K. (1996) 'Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces' *Information Systems Research*, 7(1), 111-134.

Strother, S. (2006) 'Evaluating fMRI Preprocessing Pipelines' *IEEE Engineering in Biology and Medicine* March/April 2006, pp. 27-41

Tancredi, L and Brodie, J. (2007) 'The Brain and Behavior: Limitations in the Legal Use of Functional Magnetic Resonance Imaging' *American Journal of Law & Medicine* (33)3 pp. 271-294

Toga, A. (2002) 'Neuroimage Databases: The Good, the Bad, and the Ugly' *Nature reviews. Neuroscience*. 3(4): 302-9.

Treloar, A. and Harboe-Ree, C. (2008). "Data management and the curation continuum: how the Monash experience is informing repository relationships". Proceedings of VALA 2008, Melbourne, February (in press).

Ure, J. et al (2007) *Data Integration in eHealth*: A Domain/Disease Specific Roadmap in: Proceedings of HealthGrid 2007, Geneva: IOS Press

Van Horn, J., Grethe, J., Kostelec, P; Woodward, J; Aslam, J., Rus, D., Rockmore, D.; and Gazzaniga, M. (2001) 'The Functional Magnetic Resonance Imaging Data Center

(fMRIDC): The Challenges and Rewards of Large-Scale Databasing of Neuroimaging Studies' *Philosophical Transactions: Biological Sciences*, Vol. 356, No. 1412, (Aug. 29, 2001), pp. 1323-1339.

Wellcome Trust (2007) *Policy on Data Management and Sharing* available at: http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position statements/WTX035043.htm

Wenger, E. (1998) *Communities of practice: Learning, meaning, and identity*. Cambridge University Press: Cambridge, MA.

Whalley, H., Gountouna, V., Hall, J., McIntosh, A., Whyte, M., Simonotto, E., Job, D., Owens, D., Johnstone, E., and Lawrie, S. (2007) 'Correlations between fMRI activation and individual psychotic symptoms in un-medicated subjects at high genetic risk of schizophrenia' *BMC Psychiatry* 7 (61) doi:10.1186/1471-244X-7-61

Wright, M and Corti, L. (2002) 'MRC Population Data Archiving and Access Project Consultants Report' report commissioned by UK Medical Research Council Head Office, London.

Yin, R. K. *Case Study Research, Design and Methods*, (2003) 3rd ed. Sage Publications: Newbury Park CA (US):

4. Appendices

4.1 Informed Consent Form

Consent was recorded using the form below was adapted from a sample consent form provided by the ESDS and available at: www.esds.ac.uk/qualidata.

Information about the SCARP Project

The project SCARP is being conducted by a research team at the Digital Curation Centre (DCC), comprising researchers based at the Universities of Bath and Edinburgh, and the Science and Technology Facilities Council (STFC). It is funded entirely by the Joint Information Systems Committee (JISC). A contact address is given at the foot of this form.

The project aims to discover more about disciplinary approaches and attitudes to digital curation through 'immersion' in selected cases; to apply known good practice, and where possible to identify new lessons from practice in the selected discipline areas and potentially develop new good practice. Case studies will seek the views of host research team members on practices of data sharing, curation, archiving, re-use and preservation; and facilitate access to DCC expertise in these areas.

Interview consent and data processing statement

If you consent to participating in the SCARP project and to any data gathered being processed as outlined below, please print and sign your name, and date the form, in the spaces provided.

All data will be treated as personal under the 1998 Data Protection Act, and will be stored securely. Data collected may be processed manually and with the aid of computer software.

Interviews may be recorded by the research team and selected excerpts may be transcribed by an independent transcriber who has signed a confidentiality agreement with them. If your interview is transcribed a copy of the transcript will be provided, free of charge, on request.

Please indicate, by ticking the boxes in sections 1 to 3 below, which of the following options you agree with: -

1. Consent for participation
My employer and I have been informed of the purpose of project SCARP, and my employer has agreed that I may participate in the project with my informed consent. I understand that I have the right to withdraw from the project at any time by informing the researcher or my employer. Accordingly I consent ☐ do not consent ☐ to participating in the SCARP project.

2. Publication.
I understand that my words may be quoted in reports or publications made available outside the research team and the JISC. I consent to my identity being referred to by pseudonym ☐ by name ☐ in any report whether internal or external.

3. Sharing data for research and educational purposes
Transcripts of my words may ☐ may not ☐ be shared for research or educational purposes.

Please print your name: _____

Signature: _____

Date: _____

Contact address: Angus Whyte Digital Curation Centre, Appleton Tower, Crichton Street, Edinburgh EH8 9LE. Email: A.whyte@ed.ac.uk

4.2 Interview Topic Guide

This guide outlines the scope of interview topics proposed for the Edinburgh case studies. A structured set of questions included below was used for initial 'background' interviews – however the interviews were actually semi-structured, in that questions varied depending on the initial responses. Follow-up interviews lasting 40-90 minutes were more loosely structured around themes 2-5 below.

Main themes

1. Background

Interviewee's role in research team. Research team's role in organization. Disciplinary background and research experience. Overview of data management activities associated with curation, associated issues & priorities.

2. Policy enablers and barriers

What are the enablers and barriers to adopting the principles, standards and concepts advocated in UK research institutions' data policies and guidelines; and how are those policies and guidelines being informed by current research practice?

3. Stewardship practices

How do research teams develop shared practices of data curation, sharing, reuse and preservation; and to what extent may similarities and differences in these practices be explained in terms of researchers' alignment with disciplines or domains?

4. Tools and infrastructure

How are practices of data curation, sharing, reuse and preservation supported with tools and infrastructure, and how might they be better supported?

5. Preserving context

What aspects of the context in which data is created and annotated are relevant to preserving its value for future research or learning, and how may this be better supported?

Background- Institutional, Project & Disciplinary

Topics related to: Research team's role in organization. Interviewee's role in research team. Disciplinary background and research experience. Overview of data management activities associated with curation, associated issues & priorities.

1. Please describe *your role*
2. What *domain or field of study* would you identify your work with?
3. How long have you been working in this field?
4. Do you see your *discipline* as the same as your current field of study? If not, how would you distinguish between the two?
5. What are the *main aims* of the research group?
6. Please describe how you work involves people from other fields: -
   - 6.1. Members of the research group?
   - 6.2. Members of other departments or research partners?
7. What is the time-span of your current project(s) within this group?
8. Who are your main funders?
9. What kinds of activity do you associate with 'digital curation'? (note in interviewee's own words, but probe using list below)
   - ☐ Managing digital information from its point of creation
   - ☐ Promoting the re-use of and adding of value to digital information
   - ☐ Ensuring the long- term accessibility and re-usability of digital information
   - ☐ Performing archiving activities such as selection, appraisal and retention
   - ☐ Ensuring that the authenticity and integrity are maintained over time
   - ☐ Performing preservation activities such as migration or emulation
   - ☐ Maintaining hardware components to enable data to be accessed and understood over time
   - ☐ Maintaining links between digital information, annotations, and other published materials
10. What kinds of *electronic primary data* do you create and/or work with?
11. What factors have recently changed the *nature* of data you create or use (e.g., new technologies, new forms of data, automatic data capture)?
12. What kinds of *secondary data* do you work with e.g.
    - 12.1. Do you reuse data from previous studies?
      - If so, from what sources? For what purposes e.g. meta-analysis, use in teaching materials?

- If not, is it something you are intending to do in future?

12.2. Do you look for data that may be available in (e.g.) external data centres or repositories?

- If so, which? If not, is it something you are intending to do in future?

13. It is said that effective preservation of data depends on *good quality description* of what it is and how it came about (context, technical, indexing). Much of this is best provided by the data originator at the time it is created. What helps you to do this for your data? What barriers are there to doing it?

14. What online resources do you use to find *relevant studies or other literature*? E.g. Bibliographic sources? Websites? Email exchanges with personal contacts? Email lists and newsgroups?

15. What *policies and standards* relating to data management, data sharing or preservation does the research here have to comply with?

16. Overall, what factors affect your need to curate and preserve research data?

(prompt from list)

- ☐ Regulatory compliance
- ☐ Statutory Compliance
- ☐ Educational / Research value
- ☐ Business/ Institutional requirements
- ☐ Risk Management
- ☐ Evidential Value
- ☐ Historical value
- ☐ Administrative value

17. What do you see as the *main challenges* to improving how primary data is managed for current and future needs?

(for each 'challenge' identified, ask…)

17.1. What happens now that needs to change, and how is that being addressed?

17.2. Who is involved in addressing the situation?

17.3. Who would benefit or be otherwise affected (stakeholders)?

17.4. What is driving change, or helping changes to go ahead?

What are the main barriers if any?

4.3 Interview Metadata

Metadata were inserted as document headers on interview transcripts as in the example below, and held in a separate table. Anonymised initials were linked to names and roles in a second table.

When: date	1.12.2007		
Who: inits - anon	HT		
SCARP inits	AW		
What: topic, purpose	Initial interview		
Where	*Place desc.*	*Location*	
	DCC	8th floor meeting room	
How	*Doc's etc exchanged*	*Source*	*Author*
	Data policy	HT	KE
	Consent form	AW	AW
Anonymised? (y/n)	Yes		
Version	Key points ☐ Edited transcript ☐ Full transcript ☐		

4.4 Questionnaire and Responses

The following questionnaire was used to inform the risk analysis Questionnaire text and responses are reproduced from the online survey tool Surveymonkey.com.

Your responses to this questionnaire will help to assess risks to data that could be addressed through funding to develop the Group's data management capabilities. It should only take about 10 minutes to answer the questions, and you are not identified personally with your responses. A summary of them for the group as a whole will be circulated and approved extracts may be included in a case study report. Your help is much appreciated!

Q1. Firstly, from your experience of working with the group and using the available information systems and support, what would you say is the likelihood of the events below?

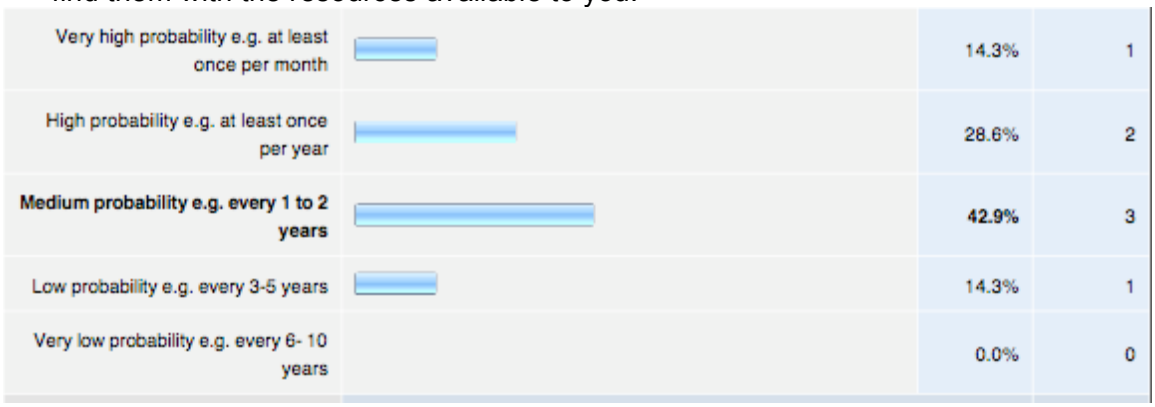
1. You need to check exactly what image processing and statistical analysis steps led to results that you wrote up, and you are able to do that from memory and the files and records kept.

Very high probability e.g. at least once per month	14.3%	1
High probability e.g. at least once per year	**71.4%**	5
Medium probability e.g. every 1 to 2 years	14.3%	1
Low probability e.g. every 3-5 years	0.0%	0
Very low probability e.g. every 6- 10 years	0.0%	0

2. You need to check exactly what image processing and statistical analysis steps you carried out to get results you wrote up, but cannot because there is no record of the script versions that were used with which data, or the steps that were taken.

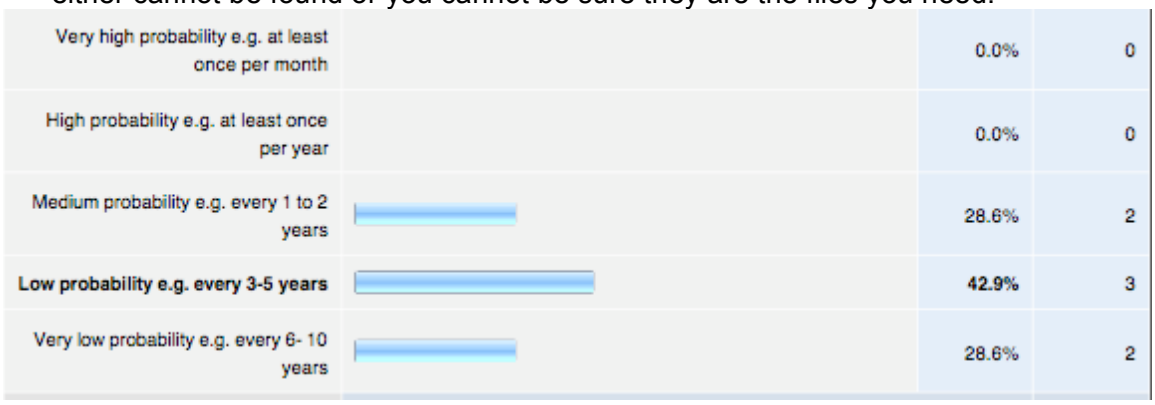
Very high probability e.g. at least once per month	0.0%	0
High probability e.g. at least once per year	0.0%	0
Medium probability e.g. every 1 to 2 years	14.3%	1
Low probability e.g. every 3-5 years	**42.9%**	3
Very low probability e.g. every 6- 10 years	**42.9%**	3

3. You need to find scans acquired for a study more than two years ago, and you can find them with the resources available to you.



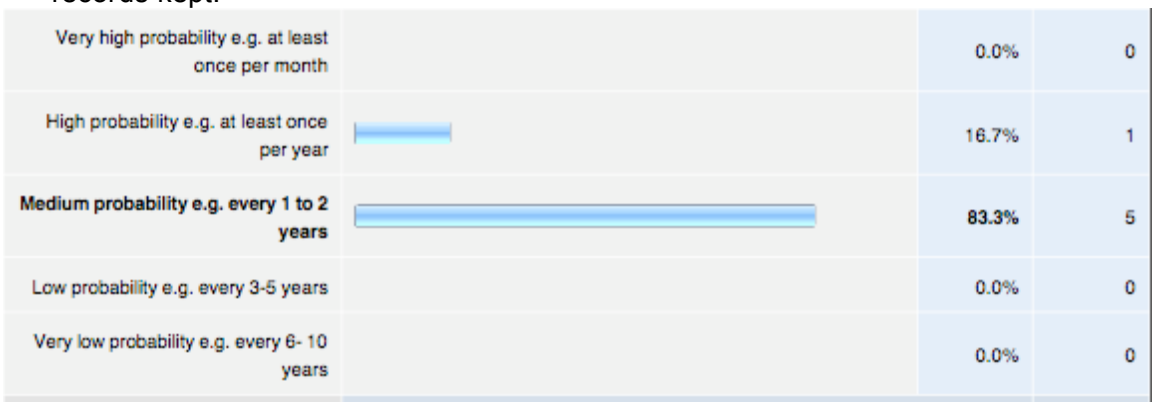
Very high probability e.g. at least once per month	14.3%	1
High probability e.g. at least once per year	28.6%	2
Medium probability e.g. every 1 to 2 years	**42.9%**	3
Low probability e.g. every 3-5 years	14.3%	1
Very low probability e.g. every 6- 10 years	0.0%	0

4. You need to find scans acquired for a study more than two years ago, but the files either cannot be found or you cannot be sure they are the files you need.



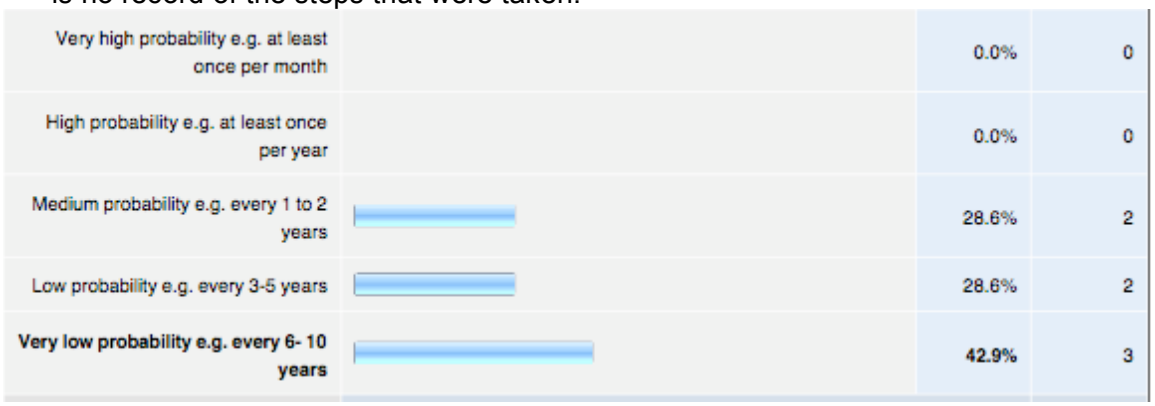
Very high probability e.g. at least once per month	0.0%	0
High probability e.g. at least once per year	0.0%	0
Medium probability e.g. every 1 to 2 years	28.6%	2
Low probability e.g. every 3-5 years	**42.9%**	3
Very low probability e.g. every 6- 10 years	28.6%	2

5. You need to check what statistical analysis steps were carried out to get results for a study more than two years ago, and you are able to do that from the files and records kept.



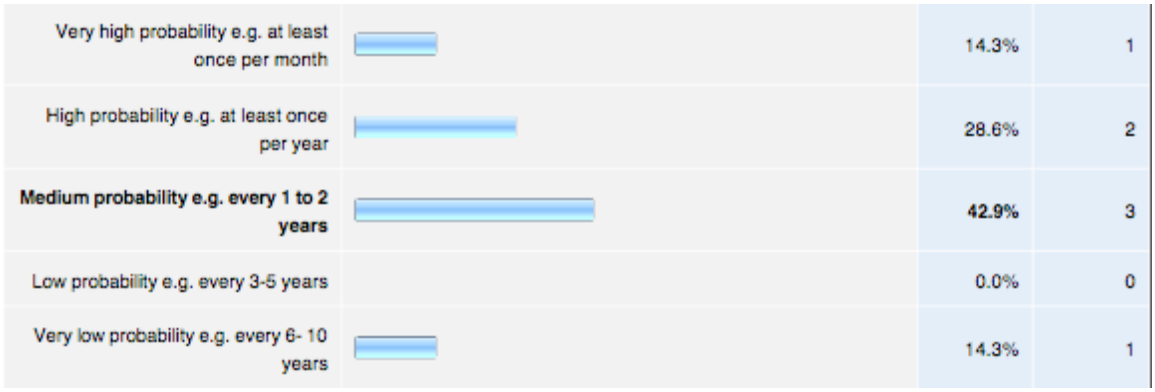
Very high probability e.g. at least once per month	0.0%	0
High probability e.g. at least once per year	16.7%	1
Medium probability e.g. every 1 to 2 years	**83.3%**	5
Low probability e.g. every 3-5 years	0.0%	0
Very low probability e.g. every 6- 10 years	0.0%	0

6. You need to check what statistical analysis steps were carried out to get results for a study more than two years ago, but cannot because the files are unavailable or there is no record of the steps that were taken.



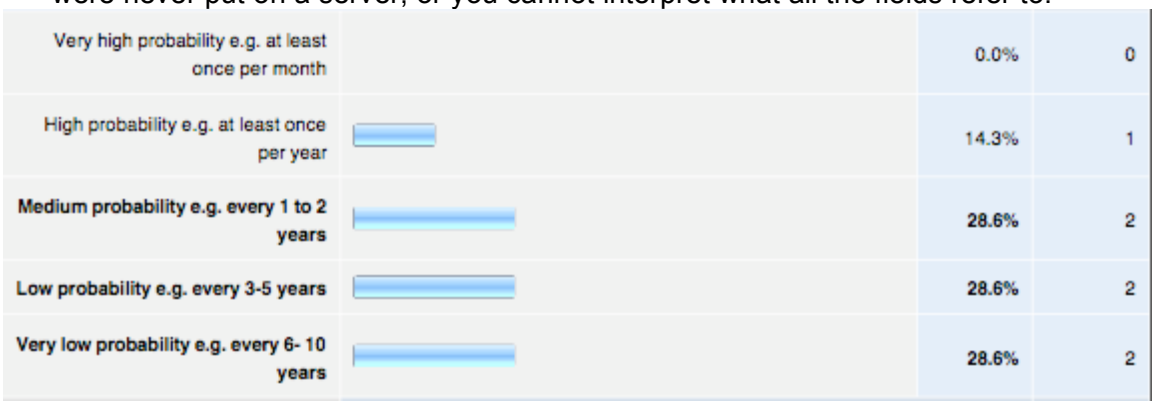
Very high probability e.g. at least once per month	0.0%	0
High probability e.g. at least once per year	0.0%	0
Medium probability e.g. every 1 to 2 years	28.6%	2
Low probability e.g. every 3-5 years	28.6%	2
Very low probability e.g. every 6- 10 years	**42.9%**	3

7. You need the demographic data and neuropsychological measures collected for a study more than two years ago, and you are able to find the files and interpret what all the fields refer to.



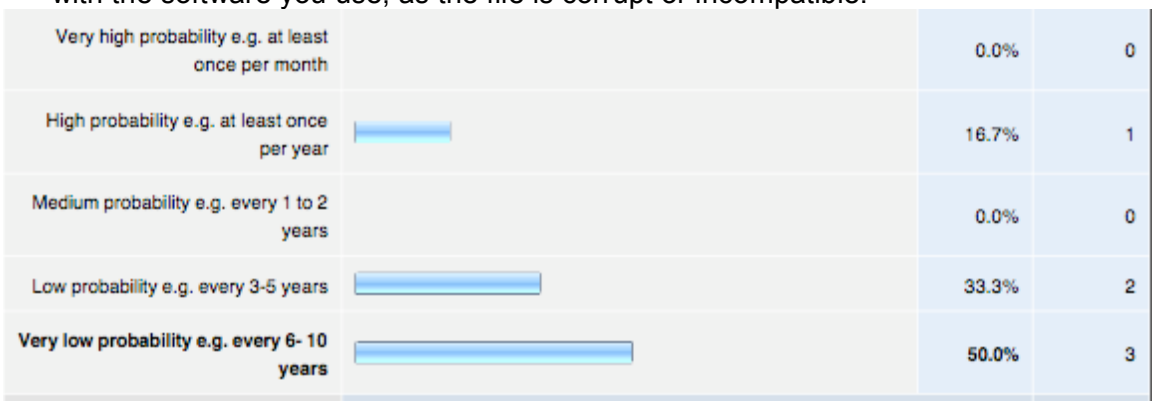
Very high probability e.g. at least once per month	14.3%	1
High probability e.g. at least once per year	28.6%	2
Medium probability e.g. every 1 to 2 years	**42.9%**	3
Low probability e.g. every 3-5 years	0.0%	0
Very low probability e.g. every 6- 10 years	14.3%	1

8. You need the demographic data and neuropsychological measures collected for a study more than two years ago- but either the files are unavailable because they were never put on a server, or you cannot interpret what all the fields refer to.



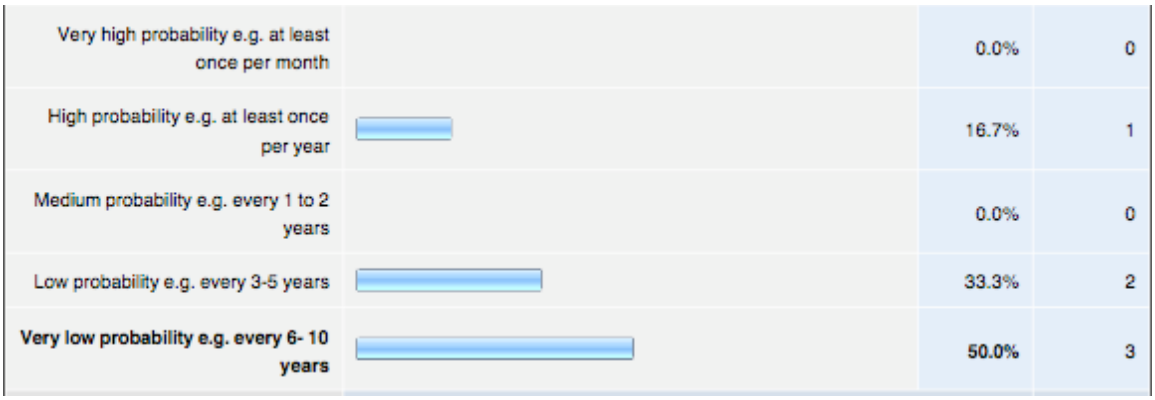
Very high probability e.g. at least once per month	0.0%	0
High probability e.g. at least once per year	14.3%	1
Medium probability e.g. every 1 to 2 years	**28.6%**	2
Low probability e.g. every 3-5 years	**28.6%**	2
Very low probability e.g. every 6- 10 years	**28.6%**	2

9. You have found a file you need, it is more than two years old and you cannot read it with the software you use, as the file is corrupt or incompatible.



Very high probability e.g. at least once per month	0.0%	0
High probability e.g. at least once per year	16.7%	1
Medium probability e.g. every 1 to 2 years	0.0%	0
Low probability e.g. every 3-5 years	33.3%	2
Very low probability e.g. every 6- 10 years	**50.0%**	3

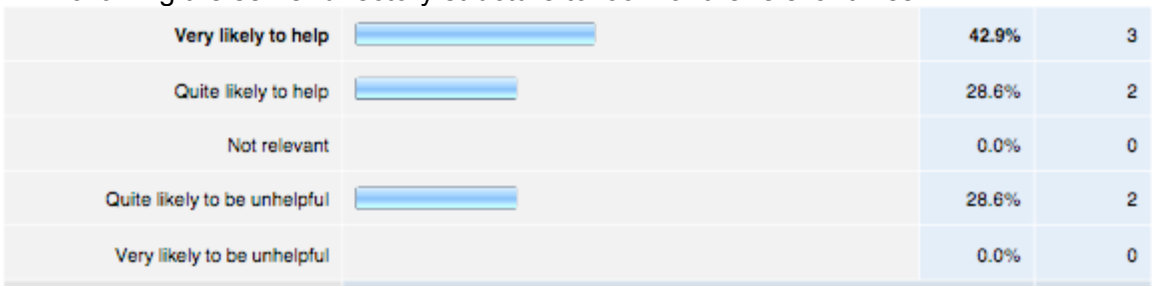
10. You need a file that is more than two years old, but the tape or disk it is stored on is unreadable.



Very high probability e.g. at least once per month	0.0%	0
High probability e.g. at least once per year	16.7%	1
Medium probability e.g. every 1 to 2 years	0.0%	0
Low probability e.g. every 3-5 years	33.3%	2
Very low probability e.g. every 6- 10 years	**50.0%**	3

Q2. You need to find and use the data from a study that is several years old and was worked on mostly by the PI and researcher who are not available to help. You have been told there is a project folder on the server that will have the scans in it, and there must be spreadsheets somewhere with the demographics, clinical and behavioural tests that were carried out. You want to bring everything together, and repeat the analysis that was done but with some new variables. How helpful would you expect to find each of the following?

1. Following the server directory structure to look for the relevant files



Very likely to help	**42.9%**	3
Quite likely to help	28.6%	2
Not relevant	0.0%	0
Quite likely to be unhelpful	28.6%	2
Very likely to be unhelpful	0.0%	0

2. Referring to a 'readme' file in the relevant project directory on the server

Very likely to help	**42.9%**	**3**
Quite likely to help	**42.9%**	**3**
Not relevant	14.3%	1
Quite likely to be unhelpful	0.0%	0
Very likely to be unhelpful	0.0%	0

3. Referring to any publications from the study and reading the methods section

Very likely to help	**42.9%**	**3**
Quite likely to help	**42.9%**	**3**
Not relevant	0.0%	0
Quite likely to be unhelpful	0.0%	0
Very likely to be unhelpful	14.3%	1

4. Raising it at a weekly meeting to see if anyone knows more

Very likely to help	**71.4%**	**5**
Quite likely to help	28.6%	2
Not relevant	0.0%	0
Quite likely to be unhelpful	0.0%	0
Very likely to be unhelpful	0.0%	0

5. Asking colleagues to forward any relevant emails exchanged at the time

Very likely to help	**42.9%**	**3**
Quite likely to help	14.3%	1
Not relevant	28.6%	2
Quite likely to be unhelpful	14.3%	1
Very likely to be unhelpful	0.0%	0

6. Asking the systems/data manager to find any files you cannot find on the server yourself

Very likely to help	**57.1%**	**4**
Quite likely to help	42.9%	3
Not relevant	0.0%	0
Quite likely to be unhelpful	0.0%	0
Very likely to be unhelpful	0.0%	0

Q.3 Information on the 'provenance' of your data and analysis would help other researchers with retrospective or secondary analysis, possibly 10-20 years from now. Provenance information includes details of what processing and analytical steps have been taken on what kinds of data, when and by whom, and for what purpose.

How far do you agree with each of the statements below about how it should be recorded?

1. I would find the time to record such information during a study for reference later when writing up

Agree a lot	**42.9%**	3
Agree a little	**42.9%**	3
Neither agree nor disagree	14.3%	1
Disagree a little	0.0%	0
Disagree a lot	0.0%	0
Don't know	0.0%	0

2. I would find the time to record such information after a paper has been accepted for publication

Agree a lot	0.0%	0
Agree a little	28.6%	2
Neither agree nor disagree	0.0%	0
Disagree a little	14.3%	1
Disagree a lot	**42.9%**	3
Don't know	14.3%	1

3. I would expect such information to be mostly recorded automatically by software

Agree a lot	0.0%	0
Agree a little	**85.7%**	6
Neither agree nor disagree	14.3%	1
Disagree a little	0.0%	0
Disagree a lot	0.0%	0
Don't know	0.0%	0

4. I would expect such information to help with retrospective analysis of our datasets

Agree a lot	**71.4%**	5
Agree a little	28.6%	2
Neither agree nor disagree	0.0%	0
Disagree a little	0.0%	0
Disagree a lot	0.0%	0
Don't know	0.0%	0

5. I would expect such information to be shared with collaborators in multi-centre projects

Agree a lot	42.9%	3
Agree a little	**57.1%**	4
Neither agree nor disagree	0.0%	0
Disagree a little	0.0%	0
Disagree a lot	0.0%	0
Don't know	0.0%	0

6. I would expect such information to be published on a website for any researcher to use, provided it does not include scans or any personally identifying data.

Agree a lot	0.0%	0
Agree a little	**85.7%**	6
Neither agree nor disagree	0.0%	0
Disagree a little	14.3%	1
Disagree a lot	0.0%	0
Don't know	0.0%	0

Finally, which of these disciplinary areas is nearest to your background?

Clinically related e.g. psychiatry, psychology, medicine	**75.0%**	3
Informatics related e.g. software engineering, computer science	25.0%	1
Other (please specify)		2